Dynamic Road Scene Classification: Combining motion with a visual vocabulary model

Anastasia Bolovinou¹², Christina Kotsiourou¹, Angelos Amditis¹
¹Institute of Communications and Computer Systems
ICCS
Athens, Greece
abolov@iccs.gr
²Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

Abstract—The majority of studies in scene classification have focused on still images, ignoring potentially informative temporal cues. This paper explores the combination of multi-scale appearance and motion features for classification of scenes captured from a moving vehicle under real-world driving. The objective is to classify unknown scenes in one out of a set of predefined typical road scene classes that are learnt during training. The method is studying the performance of a state-of-the-art scene classification visual vocabulary model (known also as bag of features model) when appearance image features and video motion features are combined for SVM learning and classification. The sequence of scenes is captured from a moving vehicle equipped with a frontal camera sensor. Video driving data used for evaluation were available by two test vehicles (a passenger car and a truck) participating in the European interactIVe IP. It is shown that a notable performance increase is realized by appearance-temporal approach in comparison to purely appearance or purely temporal methods. The quantitative evaluation has been performed using manually annotated video sequences.

Keywords—scene classification; CENTRIST descriptor; HIK clustering; motion features; multi-scale grid in space and time

I. INTRODUCTION

Natural scene classification is a fundamental challenge in the scope of automated image understanding. The ability to distinguish scenes is very useful as it can serve to provide priors for the presence of objects [1] or dynamic events (actions) [2] and even help recognize surfaces [3]. From a data mining perspective, it is a valuable tool for the management and the maintenance of large visual archives generated every day by digital visual systems worldwide [4].

Over the past decade, there has been significant research effort dedicated to the development of intelligent driver assistance systems, intended to enhance safety by monitoring the driver and the on-road environment [5, 6] by both the European commission and the automotive industry. In particular, surround analysis and understanding using vehicle-based sensing is considered crucial to enhancing the safety of drivers, vehicle occupants, and other road users [6]. In this work, we assume that the vehicle is only equipped with a frontal camera sensor and the objective is to recognize the road scene category of a single image by processing logged image sequences.

While research in scene classification has achieved considerable levels of performance based solely on local appearance [7-11], taking into account the context in a scene remains an open issue as the extraction of higher level contextual relations of visual features that can represent how a visual scene is perceived by a human is not a trivial task. In this paper we argue that still image classification can proliferate of temporal cues available from the video context. In particular we investigate whether the motion of the scene elements themselves can help discriminate among scenes with similar static appearance.

For instance, consider the two frames on the left of the arrow in figure 1. Both the frames depict a two-lane road with light traffic. However the sequence of frames (to the right of the arrows) gives further information that in the first case there is indication of moving vehicles at the far left of the scene (in the opposite direction) while the driver also approaches a bridge as we are in a highway instead of a rural road. As we are primarily interested to recognize images captured from a moving vehicle camera, these extra motion cues can be of high importance especially in cases where the field of view of the camera is restricted (e.g., due to traffic or road inclination and curve) and big parts of the image are covered by non-textured regions as road and sky which yield pure appearance information.

As studied in [12], motion information of a scene can be globally described by attributes such as:

![Fig. 1. On the left of the arrows, we show two visually similar frames (though the amount of vegetation is different), one taken from a highway road scene (up row) and one taken from a rural road scene (row below). Frames to the right of the arrow show that further information can be deducted when temporal evolution of the scene is considered.](image)
- The degree of busyness: characterizes the amount of activity happening in the video. Consider scene on an empty highway vs. scene in an urban traffic jam. The first scene is cluttered with a high degree of repeating motion patterns mainly from image parts outside the road itself, whereas in the second scene the motion patterns change much more slowly due to very low vehicle velocities. Such scene-related dynamics may be detected more reliably if a motion feature robust to camera motion due to vehicle pitch is employed.

- The degree of optical flow granularity: coarse to fine motion granularity may appear when the vehicle is driving through a flat rural road (e.g. traversing a grassland) to a rural road featured by high-vegetation lying outside the road edge. Such granularity aspects may be revealed if a multi-scale image feature extraction approach is employed.

As the above examples illustrate, global motion information can be exploited significantly to augment the spatial description of a scene. This paper shows that a notable performance increase in dynamic road scene classification is realized by a multi-scale combined appearance-temporal approach in comparison to purely appearance or purely temporal methods. To the best of the authors’ knowledge, this is the first study of motion and appearance combination for video scene classification where the images are captured from a moving vehicle and thus the camera motion highly affects the quality of the captured image.

Following the recent trend in scene recognition where local gradient-based features capturing orientation through different spatial scales of the image, yield very good results, motion boundary histograms extracted from spatio-temporal pyramids are studied as complementary cues for CENTRIST-based (CENTRIST stands for CENsus TRansform hiSTogram) [10] road scene classification. Motion boundaries [13] are selected as they are based on optical flow derivative computations which retain relative motion while motion due to locally translational camera movement is cancelled out. In order to feed a Support Vector Machine (SVM) classifier with the fused image classification vector, the powerful Bag of features model [14] is used to model the image spatio-temporal appearance.

This paper is organized as follows. Section II reviews related work. Section III presents our video representation by a vocabulary of CENTRIST and motion boundary features. The experimental setup and the datasets are described in section IV while experimental results are discussed in section V. Section VI concludes the paper.

II. RELATED WORK

Due to the fact that a camera image provides the possibility to gather a huge amount of information for many different applications it is expected that in future most of the vehicles will be equipped with a camera system. Compared to stereo systems that provide additional color or depth information, a smaller variety of potential cues is given by a monochrome camera sensor. Stereo techniques feature the advantage of explicit computation of depth and location in real-world coordinates [15] but require additional specialized hardware, precise calibration, and additional computational cost. Monocular techniques have advantages of lower cost at runtime, high temporal sampling rates and established machine learning paradigms. Therefore it seems useful to explore the viability of passive electro-optical sensing technologies such as video cameras that recently may offer very good resolution and range features coupled with low market cost.

There has been significant work in recent years to recognize natural scenes captured from digital cameras against the wide range of illumination and scale conditions that may apply in natural settings [7-11, 16, 17]. Towards this objective, one of the first successful direction established was to bypass the computationally demanding object recognition and segmentation processes and instead rely on the global layout (the “gist” of a scene) given by aggregated statistics of early visual cues, such as the power spectrum, orientation and color, e.g., [7, 19, 21, 22]. In this consideration, a forest scene presents highly textured regions (trees, grass), a mountain scene is described by an important amount of blue (sky) and white (snow), or the presence of normally distributed horizontal and vertical edges denotes an urban scene. In this case the representation consists in the extraction of ordered features of equal length measured from the image or a set of an image partitions.

Although global features enjoy the merits of low dimensionality and fast extraction, they can’t cope with scene’s uneven illumination or affine transformations often present in real-world images. Practical needs for robust object recognition under various imaging conditions and recent advances in machine learning algorithms (clustering and classification) opened the road for methods based on a dense local representation of an image.

A popular approach is the use of vector quantization on features extracted from random image patches sampled at various spatial and scale resolutions [8] of the original image to generate visual codebooks for representation and retrieval. In codebook representation, a histogram distribution of features assigned to the closest visual code-words is used as the image classification vector. This vocabulary representation based on clustering of multiple local features is also known as bag-of-features model (BoF). After the influential work of [19] on local scale-invariant descriptors, many studies have followed implementing affine invariant local descriptors that could be used for the construction of a discriminative visual vocabulary by applying appropriate clustering schemes e.g. [11, 16, 23]. The main advantage of such a vocabulary, apart from its simplicity (feature quantization is an unsupervised process and no underline data distribution is assumed), is that it inherits the invariance properties of these local descriptors.

One step further, recent studies have attempted to extend the BoF model by incorporating spatial or contextual relationships when projecting image features on the visual
words’ finite space [11, 24, 25]. Although fast and efficient techniques have been demonstrated [11, 26], inference of spatio-contextual relationships among unordered sets of visual words usually results in multi-dimensional features that require additional computational costs for feature extraction which may be prohibitive for real-time applications (note though that quantization of multi-dimensional features is not considered a problem as it is an offline process and classification can be very fast even with high dimensional classification vectors).

While the vast majority of the literature has focused on scene recognition from image stills, three notable exceptions have appeared [2, 12, 27], where histograms of optical flow (HOF) [2], chaotic system parameters [12] and spatio-temporal oriented energy features [27] are used to model scene dynamics. A drawback of the popular optical flow feature is that it is based on the assumption of brightness constancy [28] whereas the classification problem in hand (images captured from a camera installed on a moving vehicle) often includes scenes in which this assumption is violated.

Local space-time features are a successful representation for action recognition in video. Feature descriptors range from higher order derivatives, gradient information, optical flow, and brightness information [29, 2] to spatio-temporal extensions of image descriptors, such as 3D-Scale Invariant Feature Transform (3DSIFT) [30], 3D-Histogram of Oriented Gradients (HOG3D) [31], extended Speeded Up Robust Features (SURF) [20], and Local Trinary Patterns [32]. In this context, only a few approaches account for camera motion [33, 34, 35].

Since in this work we are only interested in global motion cues from scenes and not in tracking specific interest points from images (needed for object/action tracking for example), we focus our interest on exploiting the Motion Boundary Histogram (MBH) descriptor introduced by [13], which in this work is applied in image grid partitions through different time grids by varying both the image scale and time scale sampling. This is a histogram-based descriptor also and thus motion vocabulary creation and fusion with the state-of-the-art CENTRIST vocabulary [10] for scene classification is straightforward. Both descriptors can be found publicly on-line1.

1 http://lear.inrialpes.fr/people/wang/dense_trajectories and https://sites.google.com/site/wujx2001/home/libhik

III. ROAD SCENE REPRESENTATION AND CLASSIFICATION

Our scene descriptor combines an appearance descriptor, CENTRIST, extracted from a single frame of a video sequence with a motion descriptor, MBH, based on optical flow spatiotemporal derivatives extracted from a sequence of frames preceding that single referral frame.

Fig. 2 gives an overview of the feature processing steps per image (first row) or image sequence (second row). In this figure, we assume that the offline step of the two vocabularies creation, namely the BoF CENTRIST and the BoF MBH, has been preceded. The appearance-temporal model is constructed by the concatenation of these two vocabularies. Section III.A presents the descriptors and how we selected them and it is followed by Section III.B that details the image and video vocabularies extraction using spatio-temporal pyramids.

A. Structure and motion descriptors

Invariance to illumination changes has to be considered for the selection of the appropriate local features since the effect of illumination changes on the camera image may be quite strong in many cases of our dataset: e.g. due to the back-lighting situation shown in Fig. 3 or due to road shadows produced by the road-side vegetation shown in Fig. 1- second row. In particular, the back-lighting can cause an overall low contrast in the images, due to the dynamic range limitation of the digital camera sensor and for that reason we will devote a separate class to this phenomenon. On the other hand, previous empirical observations show that a scale and rotation consistency exists among scene images of the same category [e.g. 36], leading thus to the conclusion that complete rotation and scale invariance may adversely affect discriminative power of the method.

Local shape/texture histogram

Recently Wu et al. [10] introduced a new global descriptor named CENTRIST (CENsus TRansform hISTogram) which was used for place categorization. CENTRIST is based on Census Transform (CT), which compares the intensity value of a pixel with its eight neighboring pixels. If the center pixel exceeds (or is equal to) the intensity value of the neighboring pixel then a bit 1 is set, otherwise a bit 0 is set. The bit stream resulting from the eight comparisons for each individual pixel is then converted into a base-10 number (Fig. 4). Hence, each center pixel is census transformed into
a value in the range \([0, 255]\). Although it is possible to arrange the individual bits arbitrarily, we followed [10] and order the bits from top left to bottom right throughout this paper. Once all CT values are calculated for each image region, one can easily transform them into a histogram with 256 bins which results in the so-called CENTRIST descriptor. Section IIIIB provides more details on the spatial sampling followed for the extraction of the CENTRIST image descriptor.

![Fig. 3. Examples of two images with back-lighting situation.](image)

<table>
<thead>
<tr>
<th>32</th>
<th>64</th>
<th>96</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>64</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>(11010110)(_2)</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>96</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 4. Example of Census Transform value of center pixel (intensity value 64) in a 9 x 9 pixel neighborhood (figure reproduced from [10]).

Like other non-parametric local transforms for intensity values, CT is robust to illumination changes and gamma variation [10]. In addition, it has the property to retain the underlying global image structure by highlighting the discontinuities of an image, which is a very useful property as the discontinuities are the most distinctive features in an image (note that other popular features like SURF or SIFT and some gist-like features are too sensitive to light changes and do not capture the structure information).

**Motion boundary histograms**

Optical flow refers to the two-dimensional motion field in the image plane. It represents the absolute motion between two frames, which contains motion from many sources, i.e., foreground object motion and background camera motion. If camera motion is considered as a scene element motion, it may corrupt the scene classification. Various types of camera motion can be observed in real-world videos, e.g., zooming, tilting, rotation, and etc. but in our case the main type of camera motion is caused by the vehicle pitch.

For the intelligent vehicle application that we examine, the camera position and pose are considered fixed (though the mounting position and view angle differs between the test passenger car and the truck) while the camera motion is governed by the vehicle’s motion (forward translation, rotational velocity or yaw rate) and the pitch of the vehicle (a movement vertical to the image plane governed by the road surface variations and the vehicle’s suspension system). In normal driving situation, the motion can be considered as locally translational and varies smoothly across the image plane and with this assumption it can be compensated.

In this work, we employ the MBH descriptor implemented in [37] as motion descriptor to encode relative scene dynamics change from frame to frame. The MBH descriptor separates optical flow into its horizontal and vertical components. Spatial derivatives are computed for each of them and orientation information is quantized into histograms. The magnitude is used for weighting. We obtain a 8-bin histogram for each component (i.e., MBH\(_x\) and MBH\(_y\)). Both histogram vectors are unit-normalized separately with their L\(_2\) norm (vector magnitude). Since MBH represents the gradient of the optical flow, locally constant camera motion is removed while information about changes in the flow field (i.e., motion boundaries) is kept.

**B. Visual vocabulary extraction from spatio-temporal pyramids**

**CENTRIST (static) sampling:** Dense multi-scale feature sampling from each 2-dimensional spatial cell is based on local rectangular patches derived if we use 16x16 image patches and densely sample features over a grid with a spacing of 2, 4, or 8 pixels. Following [8], we also vary the grid in each image by 5 spatial scales (the spatial scale increases by a factor of \(\sqrt{2}\)). All feature vectors are scaled and rounded such that a histogram only contains non-negative integers that approximately sum to 128 since this is a pre-condition in order to use HIK-kmeans algorithm [23].

**MBH (dynamic) sampling:** Dense multi-scale feature sampling in space and time is based on dividing sampled sequences of video data (discrete frame sub-sampling interval R is set to 10 which leads to a frame rate of 3 fps for our sensor) into a spatio-temporal grid of size \([n_x \times n_y \times n_t]\), where \(n_{x,y,t}\) define the spatial-x, spatial-y and temporal-t resolution respectively, as illustrated in Fig 5b. We compute a descriptor (e.g., MBH\(_x\), MBH\(_y\)) in each cell of the spatio-temporal grid, and the final descriptor is a concatenation of these descriptors. In order to keep account of the granularity of the motion patterns, we vary the temporal-t resolution over three values: per 3, per 6 and per 9 frames. The multi-grid parameters for our experiments are set to \([n_x \times n_y \times n_t] = [(1,2,3), [1,2,3], [3,6,9]]\). This results in a maximum of a 648-dimensional descriptor (i.e., \(3 \times 3 \times 9 \times 8\)) for both MBH\(_x\) and MBH\(_y\) leading to a 1296-dimensional MBH descriptor. We evaluate the performance of different grid parameters in section V.A. If the sampling interval is set to \(R=10\) and \(n=9\), this translates into a look back in our dynamic scene data of 3 seconds duration, or equivalently we make use of 9 frames history out of a motion record of 9 x 10 = 90 frames each time we perform scene classification. In order to keep account of the granularity of the motion patterns, we repeat the MBH extraction for 3 spatial scales of each frame similar to what we did for CENTRIST sampling.

**Visual vocabulary creation:** For static scene representation, BoF\(_{CENTRIST}\) is computed based on a set of randomly selected images from each category and the sampling strategy described above. For dynamic scene representation,
on the other hand, the space-time volumes described above are used to construct a codebook for each descriptor (i.e. MBHx, MBHy) separately. We fix the number of visual words per descriptor to 200 which has shown to empirically give good results for our datasets. To limit the complexity, we cluster a subset of 100,000 randomly selected training features using Histogram Intersection kernel-kMeans (HIK) [23] instead of the popular k-means. Descriptors are assigned to their four closest vocabulary word (4N soft-assignment) using the Histogram Intersection (HI) distance instead of the Euclidean distance by modifying the weighting described in [38]. For the BoF$_{CENTRIST}$ vocabulary generation we used $K = 200$ visual words while for both BoF$_{MBHx}$ and BoF$_{MBHy}$ vocabulary generation we used again $K' = 200$ visual words. Mind that the HIK clusterer of [23] runs only on vectors of positive integer values and their magnitude cannot exceed an upper bound (in our case upper bound is set to 128).

**Scene representation:** The use of spatial and spatio-temporal grids is again employed in this step in order to incorporate spatial and spatio-temporal information. The resulting histograms of CENTRIST, MBH, and MBHy word occurrences per each spatial or spatio-temporal cell are concatenated and used as video representations. In particular, for CENTRIST image representation, we use a grid pyramid image division of 31 spatial sectors (the image is resized between different levels so that all blocks contain the same number of pixels) as proposed by [10] and illustrated in Fig. 5a, resulting in a 6200-dimensional image histogram. For MBH image representation we use a spatio-temporal grid of $\{3 \times 3 \times 9\}$ cells which was showed to perform best when cross validating on our training set, resulting in a 16200-dimensional video histogram.

**C. Classification with RBF-Chi-sq and HI kernel**

Given normalized histograms $h_i(k)$ and $h_j(k)$ with K bins we compare two non-linear kernels for SVM classification:

i) the RBF kernel with $X'$ (Chi-sq) exponential decay (simple and effective means of comparing two histograms, its properties are derived from statistics, see [39]):

$$K_{d,RBF}(x,y) = e^{-d_j}$$

$$d_j = \frac{1}{2} \sum_{k=1}^{K} \frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)}$$

ii) The histogram intersection kernel (attractive due to its ability to handle partial matches when the areas of the two histograms -the sum over all the bins- are different [40]):

$$K_{HI} = \sum_{k=1}^{K} \min(h_i(k), h_j(k))$$

Note that, recent theoretical advances on kernel SVM classification, place HI kernel classifier in the same order of computational cost for evaluation as linear SVM [40], while methods using this non linear kernel have demonstrated state of art results for scene [23] and object classification [40]. The one-against all technique was used for the multi-class classification problems in hand. For the implementation of the SVM classifier, the public LIBSVM library [41] is used.

**IV. EXPERIMENTAL SETUP AND DATASET**

In the framework of the interactIVe European interactIVe IP [6], part of which was the development of an in-vehicle perception platform, the video driving data shared among the development teams of the project had to be properly selected and categorized by the authors in order to create an annotated video/image dataset ready for classification. Data from two of the project test vehicles (a passenger car and a truck) that have been driven under various illumination and weather conditions on urban and suburban roads of Gothenburg and Aachen cities (by the Volvo Technology and the Ford Research Center in Aachen teams respectively), have been selected for that purpose (copyright of the dataset belongs to the interactIVe project).

The main role of a scene classifier for Advanced Driver Assistance Systems applications is primarily to discriminate among highway, rural and urban setting (assuming that no digital map information is available) and secondarily to recognize adverse driving settings as specifically driving in snow, presence of back-lighting (described in sec. IIIA) and driving inside a tunnel (detection of this latter case was based on the possibility of a radar processing module to tune its method in order to filter out numerous radar detections generated from the tunnel walls’ reflections). The above considerations led us to the following splitting of the dataset into 7 semantic classes (an example of each class is shown in Fig. 6) each containing 10 to 20 videos of 2-3min duration:

- **Highway** (split in two): scene of a 2 directional road (2-3 lanes per each direction) usually with a physical divider in the middle containing many road signs and road exits. This class is further divided into “highway-smooth” and “highway-traffic” accounting for the traffic density condition around the ego-vehicle.
- **Rural**: scene of a one or two directional road (usually one lane per each direction) with no physical divider in the middle and usually with the presence of off-road vegetation.
- **Urban**: scenes of one or two directional road surrounded by city visual patterns (buildings, pedestrians, bicyclists, intersections, trees).
- **Snow**: Mixed rural/highway dataset with presence of snow on non-road parts.
- **Back-lighting**: Presence of strong sun reflections crosswise the driving direction.
- **Tunnel**: Driving inside a tunnel. Presence of horizontal lights’ sequence and adjacent walls.

The capturing device is an automotive grade monochrome high dynamic range (HDR ~ 120dB) camera with wide 752x480 resolution and about 40° horizontal field of view optics. The sensor’s frame rate is 30fps which is quite high for our application so a sub-sampling of a factor...
V. EXPERIMENTAL RESULTS

This section evaluates our CENTRIST and MBH descriptors for the task of scene recognition on the seven classes described before. The descriptors and SVM kernels employed for classification are evaluated in section V.A while the computational complexity is analyzed in section V.B. Performance is measured as the mean classification accuracy on all categories (i.e. average of the diagonal entries in the confusion matrix) in a 5-fold cross-validation setting. Unless stated otherwise, we set n_x to 3, n_y to 3 and fix the motion history length, n_t, to 9 frames.

A. Comparison of different descriptors and SVM kernels

We report the performance of each individual descriptor (i.e. BoF_CENTRIST, BoF_MBH_x,y) and of the overall combination using vector concatenation and feeding the fused image vector to the SVM classifier. The results for MBH descriptor is similarly obtained by concatenating BoF_MBH_x and BoF_MBH_y image representations.

<table>
<thead>
<tr>
<th>Scene Clases</th>
<th>Mean Performance (%) per scene class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Static (CENTRIST)</td>
</tr>
<tr>
<td>highway -smooth</td>
<td>83.6</td>
</tr>
<tr>
<td>highway -traffic</td>
<td>82.4</td>
</tr>
<tr>
<td>rural</td>
<td>73.3</td>
</tr>
<tr>
<td>urban</td>
<td>85.2</td>
</tr>
<tr>
<td>snow</td>
<td>71.2</td>
</tr>
<tr>
<td>back-lighting</td>
<td>72.3</td>
</tr>
<tr>
<td>tunnel</td>
<td>84.1</td>
</tr>
<tr>
<td>Avg (%)</td>
<td>78.9</td>
</tr>
</tbody>
</table>

As it is shown in Table I, BoF_CENTRIST is a powerful appearance descriptor that outperforms motions descriptors showing that spatio-appearance information can be used for discriminating scene classes for images captured from a moving vehicle in real-world driving scenarios (capturing of structured patterns like buildings explain the high recognition rate in urban class). Motion descriptors on the other hand, although their individual performance is lower than CENTRIST, provide complementary information leading to an average increase of performance of 2.5%.

MBH_y outperforms MBH_x by 3% in all classes but the “tunnel” class. This is presumably due to the fact that scene movements in the road scenarios (e.g. see fig. 6a-f where static scenery on the left and on the right of the road being traversed and optionally oncoming vehicles describe the scenes) include motion patterns which are dominant in the vertical direction (because 3D traffic in a straight road towards the z-axis in real-world coordinates is translated in 2D vertical movement). The opposite effect for the “tunnel” class may be explained by the various horizontal light reflections (see fig 6.g) that are present in these images (wall lights) and are captured by the camera sensor.

The effect of the spatio-temporal grid resolution is studied in Table II where the mean performance of BoF_MBH_x,y versus 6 configurations of the \([n_x \times n_y \times n_t]\) grid is reported. As suspected, global motion attributes are better represented by a longer video history (9 sub-sampled frames every 3secs) in the denser spatial grid (3 x 3) and thus a trade-off between increased spatio-temporal resolution and desired dimensionality of the motion descriptor should be achieved. Moreover, frame history longer than some frames can hurt the performance, as the video extract has a higher chance to belong in more than one class. In our case, the transitions among scene classes in videos are not so rapid and thus we don’t have to worry about that.

<table>
<thead>
<tr>
<th>([n_x \times n_y \times n_t]) grid</th>
<th>Mean Performance over all classes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x1x3 (1 sec history)</td>
<td>34.9</td>
</tr>
<tr>
<td>1x1x6 (2 secs history)</td>
<td>38.2</td>
</tr>
<tr>
<td>1x1x9 (3 secs history)</td>
<td>51.2</td>
</tr>
<tr>
<td>3x3x3 (1 sec history)</td>
<td>46.7</td>
</tr>
<tr>
<td>3x3x6 (2 secs history)</td>
<td>49.3</td>
</tr>
<tr>
<td>3x3x9 (3 secs history)</td>
<td>64.2</td>
</tr>
</tbody>
</table>

B. Comparison of different descriptors and SVM kernels

<table>
<thead>
<tr>
<th>SVM kernel</th>
<th>Mean Performance (%) over entire dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static (CENTRIST)</td>
<td>Dynamics</td>
</tr>
<tr>
<td>RBF-Chi_sq</td>
<td>75.4</td>
</tr>
<tr>
<td>HI</td>
<td>78.9</td>
</tr>
</tbody>
</table>

As shown in Tables III, in the case where dimensionality of the classification vector increases from 6200 to 6200+ \(\{2x16200\} \cdot d = \{38600\} \cdot d\), HI kernel consistently outperforms RBF-Chi-sq kernel confirming that HI kernel is the most appropriate kernel when dealing with high dimensional histograms.
C. Computational effort for training/testing

All routines are implemented in C++ using VS.Net platform. C++ executables ran on a 4G Ram Pentium 3 with 4 processing units. As Table IV shows, the offline procedure of clustering visual and motion descriptor data into 200 visual words with the HI-kernel k-means algorithm is quite efficient, e.g. approximately 9k histograms from 35 training videos are clustered across the training set (incl. the seven classes) in about 2 minutes. The time spent for classifying an unseen image represented by a vocabulary-based vector of ~7500-d (i.e. BoF\textsubscript{CENTRIST}, + BoF\textsubscript{MBHy}, + BoF\textsubscript{MBHx}) was around 1.95 seconds per image which is still below the 3 seconds period of the input data history (see sec. IIIIB) and can be tolerable for monitoring applications with non real-time constrains.

Note that the speed can be twice faster when features are only extracted on the original scale. The video sampling rate can be tolerable for monitoring applications with non real-time constrains.

TABLE IV. COMPUTATION TIMES SPENT ON THE MAJOR STEPS OF THE METHOD

<table>
<thead>
<tr>
<th>Total time (secs)</th>
<th>Percentages of time spent during training</th>
<th>Opt. Flow</th>
<th>CENTRIST</th>
<th>MBHy</th>
<th>MBHx</th>
<th>Save features</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>428.6</td>
<td></td>
<td>29%</td>
<td>9%</td>
<td>19%</td>
<td>15%</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td>1.95</td>
<td>Descriptors extraction and assignment</td>
<td>85%</td>
<td>Classification</td>
<td>15%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

This paper introduced an approach for efficient scene recognition from a moving vehicle based on fusion of CENTRIST and motion boundary histogram descriptors extracted through a multi-scale grid in space and time. Our fused dynamic appearance-temporal representation was shown to outperform the baseline static appearance representation. Future work could improve the efficiency of the motion feature extraction by ego-motion compensation or parallelization techniques or representation with more efficient features (e.g. 4C) of the proposed method one would need to replace GPU/CPU implementation.

Note that the speed can be twice faster when features are only extracted on the original scale. The video sampling rate can be tolerable for monitoring applications with non real-time constrains.

ACKNOWLEDGMENT

This work was also supported by the European Commission under interactIVE, a large scale integrated project part of the FP7-ICT for Safety and Energy Efficiency in Mobility. The authors would like to thank all partners within interactIVE for their cooperation and valuable contribution and especially the VTEC and FFA vehicle demonstrator teams who logged the training/test data.

REFERENCES


